Using LaTeX to Typeset a Marāṭhī-English Dictionary

Manasi Athale and Rahul Athale
Research Institute for Symbolic Computation
Johannes Kepler University
Linz
Austria
manasi,athale@risc.uni-linz.ac.at

Abstract

We are using LaTeX to typeset an old Marāṭhī-English dictionary, dated 1857. Marāṭhī is the official language of Mahārāshtra, a western state of India. Marāṭhī (मराठी) is written using the Devanāgarī script. The printed edition of the dictionary contains approximately 1000 Royal Quarto size ($9\frac{1}{2}'' \times 12\frac{2}{3}''$) pages with around 60,000 words. The roots of the words come from many languages including Sanskrit, Arabic and Persian. Therefore the original dictionary contains at least *three* different scripts along with many esoteric punctuation marks and symbols that are not used nowadays.

We have finished typesetting 100 pages of the original dictionary. We present our experiences in typesetting this long work involving Devanāgarī and Roman script. For typesetting in Devanāgarī script we used the `devnag` package. We have not yet added the roots in other scripts but that extension can be achieved with the help of ArabTeX. We want to publish the dictionary in electronic format, so we generated output in PDF format using pdfLaTeX. The bookmarks and cross-references make navigation easy. In the future it would be possible to design the old punctuation marks and symbols with the help of METAFONT.

1 Introduction

Marāṭhī is a language spoken in the Western part of India, and it is the official language of Mahārāshtra state. It is the mother tongue of more than 50 million people. It is written in the Devanāgarī script, which is also used for writing Hindi, the national language of India, and Sanskrit. The script is written from left to right. A consonant and vowel are combined together to get a syllable, in some cases consonants can be combined together to get conjuncts or ligatures. While combining the vowel and a consonant one might have to go to the left of the current character — which is a big problem for a typesetting program.

We are typesetting a Marāṭhī-English dictionary compiled by J. T. Molesworth and published in 1857. The dictionary is old so there is no problem about copyright. This will be the first Marāṭhī-English dictionary in an electronic format.

2 Devanāgarī Script

There are 34 consonants, 12 vowels, and 2 vowel-like sounds in Marāṭhī. Table 1 gives the consonants along with some common English words to illustrate the sounds. In some cases, there is no exact equivalent English sound, and we give those with standard philological transliteration. The *h* in this table designates aspiration, and a dot under a consonant designates retroflexion. Although Hindi and Marāṭhī use the same Devanāgarī script, the consonant ळ, which is used in Marāṭhī is not used in Hindi. Similarly some characters used in Sanskrit are not used in Marāṭhī. All the consonants have one inherent vowel अ (*a*), and in order to write the consonant itself without the vowel, a special "cancellation" character (्) called *virāma*, must be used. For example, स, is स् + अ, where अ is a vowel.

Table 2 lists the vowels and the two vowel-like sounds. The first two rows give the vowels and the last row gives the vowel-like sounds, called *anuswāra* and *visarga*, respectively. In general, the vowels are paired with one member short, the other long. For ऋ, make the *r* into a syllable by itself.

A vowel is added to a consonant to produce a syllable, for example all the consonants written above already have the vowel अ. Suppose we want to get the sound, *Sa*rah, with a long *a*. We add the second vowel आ to स् to get सा, where we can see a bar added behind the consonant स.

क	ख	ग	घ	ङ
*c*ar	*kh*	*g*o	*gh*	nasal
च	छ	ज	झ	ञ
*ch*air	*cch*	*j*ail	*z*ebra	nasal
ट	ठ	ड	ढ	ण
ṭ	*ṭh*	*ḍ*	*ḍh*	*ṇ*
त	थ	द	ध	न
*T*ehran	*th*	*d*ark	*dh*	*n*ew
प	फ	ब	भ	म
*p*air	*f*ail	*b*at	*bh*	*m*an
य	र	ल	व	
*y*ellow	*r*oad	*l*ove	*w*ay	
श	ष	स	ह	ळ
*sh*are	*ṣ*	*s*un	*h*appy	

Table 1: Devanāgarī consonants.

अ	आ	इ	ई	उ	ऊ
*a*bout	c*a*r	s*i*t	s*ea*t	p*u*t	r*oo*t
ऋ	ऌ	ए	ऐ	ओ	औ
und*er*	bott*le*	s*ay*	b*y*	r*oa*d	l*ou*d
अं	अः				

Table 2: Devanāgarī vowels.

Now to write *si*t we add the third vowel इ to स्, and here it gets difficult, because although the *i* is pronounced after the *s*, it is written before the consonant. We get सि, where a bar is added before the character स.

Syllables can even be formed using more than one consonant and a vowel. For example, द्‌वितीया, here we add द्, व् and इ. It can also be written as द्वितीया. There are many such variations when two or more consonants are combined, and some conjunct characters look nothing like their constituent parts. For example, र or *r* is written in *four* different ways depending on the other consonant in the conjunction.

The first vowel-like sound, *anuswāra*, is the nasal consonant at the end of each of the first five consonant rows in the consonant table. For example, गंगा (Ganges), here the "ं" on the first character is the *anuswāra* but it is pronounced as the nasal sound in the row of the next character, which is गा. The sound is like ङ. *Visarga* is more or less a very brief aspiration following the inherent vowel (and for this reason it is usually written *ḥ* in philological transcription).

3 Problems

We tried many approaches before choosing LaTeX. Scanning the pages was out of question as the printed quality is very poor. Also many of the consonant conjuncts are written in a different way nowadays, so it would be difficult for the average modern reader to decipher the old dictionary. There are some web sites that have dictionaries in two scripts using Unicode. But in many cases it does not show the correct output, and it is difficult to find suitable viewers. We thank referees for mentioning an XML approach, but we did not try that. We also tried Omega, but there was hardly any information available when we started our work more than two years ago and also the setup was very difficult.

The first problem was having two scripts in the text, and typesetting it such that both scripts mesh well. Molesworth uses Marāṭhī words to explain the concepts so Devanāgarī script text appears also in the meaning. Also there are couplets of a poem in places to explain the usage. Many Marāṭhī words have roots in Sanskrit, Hindusthani, Arabic and Persian. Arabic, Persian, and the Urdu variant of Hindusthani are written using the Arabic script, which is the third script used in the dictionary. In Marāṭhī, a word is spoken—and also written—in a slightly different way depending on the region of the publication. Therefore in the dictionary, the most used form usually has the meaning listed for it, and all other forms have a note pointing to the most used form. This requires cross-referencing for faster use.

The dictionary has a long preface giving the details of how the words were chosen, which meanings were added, and so on. It contains different symbols and punctuation marks. Also in the meaning of some words, symbols are used to show the short form used during that period, which is obsolete now.

The printed dictionary is heavy, so carrying it everywhere is out of question. We wanted to give the user the possibility to carry the dictionary on a compact disc or computer. Therefore the next question was, which is the most user-friendly and/or popular output format?

4 Solution

In a single word: pdfLaTeX. We mainly used two packages to do the typesetting: `lexikon` for dictionary style, and `devnag` for Devanāgarī script. It is a two step process to typeset in Devanāgarī script. A file, usually with extension `.dn`, is processed with `devnag`, a program written in the C language, to get a `.tex` file. The preprocessing step is necessary due to the problem of vowel placement, complex conjunct characters, and so on, as mentioned in the introduction. The style file `dev` is used to get the Devanāgarī characters in the output PDF file after compiling using pdfLaTeX.

Once we have a `.tex` file we can get output in many formats, DVI, PS, PDF, etc. We chose PDF as there are free readers for almost all platforms and pdfLATEX makes it easy to go from TEX to PDF. The `hyperref` package solved the problem of cross-referencing and bookmarks. The user can click on a hyperlinked word to go to the form of the word that has the complete meaning, and come back to the original word with the back button in his favourite reader. In addition to the hyperlinks, bookmarks make navigation much easier; for example, bookmarks point to the first words starting with *aa*, *ab*, *ac*, etc. An additional nested level of bookmarks is chosen if there are many words starting with the character combination. For example, if there are many words starting with *ac* then we also have bookmarks for *aca*, *acc* and so on. Usually there are fewer than five pages between two bookmarks, so finding a word is not time consuming.

The preface contains characters like आा, which is not part of the modern Marāṭhī character set, but which was used as a short form a hundred years ago. To typeset this character we directly edited the `.tex` file after preprocessing to get the required result.

We have attached at the end of this article an annotated sample page from the typeset dictionary. At the top of the page the first entry is the first word on the page, then the copyright information with our name for the dictionary, शब्दविश्व, simply translated as “the world of words”, followed by the entry of the last word on the page. On the right hand side is the page number. At the bottom, the page number is given in Devanāgarī script.

5 Future Work

After completing the typesetting the whole dictionary we will add the roots of the words in Hindusthani, Arabic and Persian. Currently we denote this using [H], [A] or [P], respectively. We have tried typesetting in three secripts on some small examples and did not find any conflicts between ArabTEX and `devnag`. We have not yet created new symbols but it is possible with the help of the `pstricks` package or METAFONT.

References

[1] Devanāgarī for TEX, `http://www.ctan.org/tex-archive/language/devanagari/velthuis/doc/devnag/manual.ps`

अकढा or अकढी *a* [P] (अ & कढणें) Unscalded or unheated—milk, oil, ghee. 2 Unheated to the degree of fusion—metals.

The root of the previous word is Persian, which is denoted by [P]. In the original dictionary the root is written in the original script. The word ghee appears in Helvetica font to stress the fact that this word is of Indian origin.

अकण *a* Devoid of कण or grit—cleaned rice. 2 wanting corn in the car—standing or thrashed crops. 3 Having no corn to eat. Ex. अकणा कण होय अधना धन होय.

अंकणकडवे *n* (See आंकणकडवे) The burden or bob of a song.

The popular form of the previous word is different, which appears on the right hand side in the paranthesis. A user can click on that word to get more meanings on the current word.

अंकणी *f* A ruler. 2 (Verbal of अंकणें) Marking &c. 3 A division (as in a box), a compartment.

The previous word is a verbal of the word in the paranthesis on the right hand side. To get the meaning of the original verb user can click on that on the right hand side.

अंकणें *v c* अंकन [S] To mark, gen.; to number, stamp, dot, rule, mark with lines, figures &c.: also to rule (lines); to trace (outlines); to delineate, sketch, roughdraw, draw, describe.

The root of the previous word is Sanskrit, which is denoted by [S]. Note that Sanskrit root is given only if it is different from the word.

अंकणें *n* The sloping divisions upon a धावे or flat earthroof for the water to roll off.

अकथित *a* [S] Untold, unnarrated, unsaid.

अकथ्य & अकथनीय *a* [S] Unspeakable, ineffable, indescribable, inenarrable. Ex. परब्रह्म कीर अ°

अंकन *n* [S] Marking gen.; numbering, stamping, dotting &c.

अंकनीय *a* [S] To be marked, numbered &c. See the noun अंकन.

अकपट corruptly, अकपटी *a* [S] Free from malice or guile; forgiving, frank, ingenuous. 2 Real, true, genuine. 3 Used as *s n* Candor, openness, ingenuousness, guilelessness, absence of malice or grudge. Ex. मी तुम्हाशी अकपटाने वागतो.

अंकपट्टी *f* The ticket or label appended (to a bale of cloth &c.) showing the number and price.

अंकपाश *m* [S] In arithmetic. Permutation.

अकबरशाई *a* Of the currency established by the emperor Akbar—a rupee &c.; relating to the reign of Akbar.